Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2714–2720
Marseille, 20-25 June 2022

Increasing CMDI's Semantic Interoperability with schema.org

Nino Meisinger, Thorsten Trippel, Claus Zinn
University of Tuebingen
Wilhelmstraße 19, 72074 Tuebingen, Germany
{firstName.lastName}@uni-tuebingen.de

Abstract

The CLARIN Concept Registry (CCR) is the common semantic ground for most CMDI-based profiles to describe language-related resources in the CLARIN universe. While the CCR supports semantic interoperability within this universe, it does not extend beyond it. The flexibility of CMDI, however, allows users to use other term or concept registries when defining their metadata components. In this paper, we describe our use of schema.org, a light ontology used by many parties across disciplines.

Keywords: Metadata Management, Semantic Interoperability, CMDI, schema.org

1. Introduction

In practise, the semantic interoperability of CMDI-based metadata is based upon a single, CLARIN-based, registry. When metadata practitioners define CMDI profiles and their components, they usually ground the elementary metadata fields in the CLARIN concept registry (CCR). This helps applications that make use of CMDI-based metadata, such as the Virtual Language Observatory (VLO), to provide a semantic search across millions of resources described with many different CMDI profiles. The grounding of CMDI profiles with the CCR facilitates the mapping of commonly used data categories to the search facets offered by the VLO. The overwhelming majority of data categories, however, are not discipline specific, nor do they only make sense in the CLARIN context. In this paper, we argue that a grounding of CMDI-based profiles with vocabularies that are widely used across communities increases their semantic interoperability. Where possible, CMDI practitioners should not rely on community-specific (and hard to maintain) vocabularies. Instead they should opt for using vocabulary that is understood outside the CLARIN realm. To test our approach, we have developed a tool that transforms CMDI-based metadata based on CCR terms into CMDI-based metadata based on schema.org terms. Our TALAR repository will now export the traditional XML-based CMDI data together with a JSON-LD-based CMDI that makes use of the new terminology.

2. Background

The Common Language Resources and Technology Infrastructure (CLARIN) enables research based on digital language resources by offering advanced services to discover, explore, exploit, annotate, analyse, combine or archive language data, see `clarin.eu`. At the core of this infrastructure serves a metadata framework that makes it possible to describe language-related resources (data and tools) with a rich, expressive vocabulary. Expressive metadata is key to implementing the FAIR principles (Wilkinson et al., 2016) of findability of research data, accessibility, interoperability and reusability: metadata must use well-defined terms and value schemes that are shared within the research community, and ideally understood across communities.

Existing metadata schemes by themselves do not fill the bill. Bibliographic metadata schemes such as Dublin Core (`dublincore.org/`) or MARC21 (`www.loc.gov/marc/bibliographic/`) are standardised schemas from the library world. While they are widely used, they only provide *domain-independent* metadata fields for the description of artefacts and prints. They lack the descriptive power for an adequate description of other types of research data sets such as lexical resources, text or speech corpora, experimental data, tree-banks, and other language-related resources.

2.1. CMDI, a FAIR Compliant Metadata Infrastructure

The Component Metadata Infrastructure (CMDI, ISO 24622-1 and ISO 24622-2, see also (Broeder et al., 2010)) was designed as a FAIR compliant metadata framework with built-in semantic interoperability, even before the FAIR principles were discussed in the research data community.

As the acronym suggests, CMDI is a metadata *framework* that allows the definition of metadata schemas. Each type of resource merits a description that captures its nature in an adequate manner, and each archive may have its specific needs that need to be reflected via the metadata schema(s) it employs. CMDI is a hierarchical metadata framework where a *profile* – from which an XML-based schema can be derived from – is built from smaller building blocks, *CMDI components*. A CMDI component can consist of a set of other CMDI components or so-called elements (*i.e.*, *data categories*). The data categories utilised in the metadata schemas are defined with reference to publicly available definitions. The reference is provided by a public identifier, usually an IRI, often a resolvable URI. The values for the data categories are defined as patterns, ranging from general

string values over specific types such as dates or closed vocabularies, which themselves can be defined in terms of an identifier, such as an IRI pointing to a concept defined in an ontology.

2.2. The CLARIN Concept Registry

The description of different types of resources require different terms sets, and they shall be precisely defined. To support the definition of such terms sets, the CLARIN community built the ISOcat registry (Broeder et al., 2010), an implementation of the ISO 12620:2009 standard (ISO 12620, 2009). Experiencing the tedious standardisation process that is defined and implemented in ISOcat, its successor, the CLARIN Concept Registry (CCR) (Schuurman et al., 2016; Wright et al., 2014), was implemented and the ISO standard ISO 12620:2019 was refined to target only terminological databases rather than providing data categories in a more general sense.

By and large, the CLARIN Concept Registry (CCR) has been adopted by CMDI metadata designers as *de facto* standard when making references to data category definitions.[1] Note, however, that the CMDI specification that not prescribe the use of the CCR but allows *any* term registry (or semantic registry) to define common ground across CMDI components and profiles.

2.3. Bridging the Gap towards Linked Data

CMDI induces semantic interoperability within the CLARIN world by using the CLARIN Concept Registry as common ground across profiles. CMDI-based profiles (and their derived XML-based schemas) hence usually share a large part of their vocabulary. Most CMDI-based profiles rely on the CCR only, and hence, are not connected to data sources outside of CLARIN.

The Linked Data (LD) initiatives connect data sets from many different domains by means of the Resource Description Framework (RDF, see `https://www.w3.org/TR/rdf11-concepts/`) and related technologies. A step towards sharing CMDI-based metadata with LD communities is, hence, to map CMDI metadata instances to RDF-based data (Windhouwer et al., 2017).

While conversion to RDF-based data is a necessary step towards data sharing with LD communities, it is not sufficient. For this, CMDI-based data must be linked to other data. Trippel and Zinn (2020) attach authority file data to persons involved in the creation of language resources. Person names, usually provided as strings in CMDI-based metadata, for instance, are complemented with persistent identifiers to their `viaf.org` identities. The VIAF linked data set links together integrated authority files from many national libraries worldwide, including for instance, the German National Library (see `d-nb.info/standards/elementset/gnd/`). The persistent identifiers provided by these libraries help to uniquely identify person and organisation names across language-specific name spellings or spelling variations, and are also used outside of the library world.

Using authority file information is an example for using well-defined value spaces or closed vocabularies for data categories. In (Zinn et al., 2012), it is illustrated how CCR-based data categories can be mapped to a vocabulary used by many disciplines and communities, namely, `schema.org`.

2.4. Schema.org

Prior to schema.org, the developers of the major search engines attempted a variety of semantic annotations of web pages to improve search results. In these early days, each search engine developer team used their own, proprietary approach for semantic annotation. Website designers, therefore, had to follow many different recommendations to ensure that their sites were listed in the result set of the main search engines. With the increasing number of topics to be annotated, existing annotation solutions did not scale well; also website creators were either confused by the various annotation methods, or refused to use them at all. Thus the idea of `schema.org` was born, a single, light ontology that was supported by all major search engines from the beginning. The ontology covers a wide range of topics, and its design and community support foster thematic scalability (Guha et al., 2016).

Among 10 billion sampled web pages, (Guha et al., 2016) found that in 2016 31.3% of them are using schema.org markup, an increase of 9.3% from a 2015 census. The first use case was for Google's "Rich Snippets" – advanced text excerpts displayed in the search results of Google such as ratings for companies or products, or the average cooking time for a recipe.

In the sequel, Google exploited schema.org annotations as a data source for the company's "Knowledge Graph", which now enriches Google search results with info-boxes related to the search query. The same approach has been followed by Microsoft's Bing engine.

Schema.org vocabulary can be represented in various formats such as RDFa, Microdata (an HTML5 standard to nest metadata inside web pages), and JSON-LD.[2][3]

Schema.org itself follows a hierarchical type-subtype structure consisting of two key building blocks: types and properties. Every type originates from the "Thing" type and inherits all properties from its parents. Properties themselves are used to describe a type in more detail, *e.g.*, a "Person" contains properties, such as "fam-

[1] The CMD Cloud and SMC Browser (Ďurčo and Windhouwer, 2014; Ďurčo, 2013), see also `https://clarin.oeaw.ac.at/smc-browser/index.html`, illustrate the current use of data categories and definitions for CMDI based metadata.

[2] See `https://developers.google.com/search/docs/advanced/structured-data/intro-structured-data#markup-formats-and-placement`.

[3] See `https://json-ld.org`.

ilyName" or "email". As values, properties either take URLs, strings, or other schema.org types.
The schema.org ontology has grown and matured over the past years. The healthy state of the project and its wide use and adoption across many communities make the light ontology a perfect candidate for increasing the semantic interoperability of our CMDI-based profiles beyond the CLARIN universe. For this, we need to map, and finalize replace, data categories defined in the CCR to, and with, concepts in the schema.org ontology.

3. Mapping from CCR to schema.org

Our TALAR repository at the University of Tuebingen hosts six main profiles. Each is targeted at describing one particular type of language resource: experiments, lexical resources, speech corpora, text corpora, tools, and so-called resource bundles. All our profiles share the same CMDI-based components when it comes to the description of resource aspects that are *not related* to the particular resource class, namely, the components "GeneralInfo", "Project", "Publication", "Creation", "Documentations", "Access", and "ResourceProxyListInfo". As it turns out, most of the data categories used in these components have equivalent terminology in schema.org, see Table 1. We have hence developed a system that supports the conversion of CMDI-based instances making use of CCR entries to those that now make reference to schema.org concepts. The resulting tool, called CMDI2JSONLD, furthermore supports a generic CMDI to JSON-LD transformation using the predefined identifiers from the CCR and Component Registry.[4]
To support the mapping process, we adopted the line of approach that is followed by the VLO facet mapping (Van Uytvanck et al., 2012). Since schema.org descriptions tend to be more complex in nature, the system had to be extended to handle a variety of different transformation scenarios.
A mapping between a CCR term and a schema.org concept is expressed in XML. The overall structure is depicted in Fig. 1. For each CMD profile, the corresponding schema.org type needs to be specified. Since most profiles tend to describe a dataset, the DataSet type of schema.org acts as the default for unspecified profiles. Each type in the mapping is paired with a JSON-LD `context` description. The `context` is freely modifiable to allow for more flexibility. It will be directly inserted into the resulting JSON-LD.
Finally, one can define mappings for all properties of the given type. This can be done in three ways:

1. Specifying a concept's identifier from the CCR. It is possible to specify multiple concepts.
2. Specifying XPath expressions in case no concepts can be used. XPath (Clark et al., 1999) is a query language for selecting nodes of an XML document.
3. Blacklisting profiles. If a profile is blacklisted, it will only be evaluated against the XPath expressions, not the concepts.

The last option is beneficial for cases where a concept entry of the CCR might not have a perfect match with a corresponding schema.org property or the element in question is not defined in the CCR, thus lacking an identifier. Thus, only mapping a pre-specified subtree of the CMDI using XPath expressions is the more sensible choice.
Due to the complexity of both schema.org and CMD profiles, a simple one-to-one mapping between the CCR entries and schema.org properties is not always possible. However, the system allows for more complex mappings rules:

1. The `license`[5] concept expects "a description of the licensing conditions under which the resource can be used". In the TALAR repository, this is realised by inserting the (string-based) name of the license. Schema.org however, either expects a URL or the CreativeWork type for their license property. Thus, to correctly map the concept it must be transformed to a CreativeWork type whose name property contains the license name.

   The system supports this kind of complex type mapping by using the "type" attribute:

```
<license type="CreativeWork">
  <name>
    <concept>URL/concept>
  <name>
</license
```

2. In some instances, entire CMD components map better onto schema.org types, compared to just the CCR entries. The creator property of schema.org, for example, expects either a type of Organization or of Person. The CMD profiles utilised in the TALAR repository, for example, have a Creators component, containing a list of all persons involved in creating a resource. Each person is described as another component, consisting of the firstName and lastName concept. Just specifying both these concepts in the mapping with the methods introduced so far, does not work, as the underlying tool for the transformation does not know

[4] See `https://weblicht.sfs.uni-tuebingen.de/converter/MetaDataTransformer/` for the web-based interface, and `https://github.com/SfS-ASCL/metadatatransformation` for the source code repository on GitHub.

[5] See `http://hdl.handle.net/11459/CCR_C-2457_45bbaa1a-7002-2ecd-ab9d-57a189f694a6`.

Field name (in CMDI)	Description Link to DC-based Definition	Definition in schema.org
ResourceName	A short name to identify the language resource CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5	/name
ResourceTitle	The title is the complete title of the resource without any abbreviations CCR_C-2545_d873f2ab-2a2f-29d6-a9ab-260cde57f227	/alternativeHeadline
ResourceClass	Indication of the class, *i.e.*, the type of a resource CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215	/additionalType
Version	A number that identifies the version of a metadata description a resource or a tool/web service CCR_C-2547_7883d382-b3ce-8ab4-7052-0138525a8ba1	/version
LifeCycleStatus	Indication of the status in the life cycle of a resource. CCR_C-3818_8c4aec73-1654-7565-9575-c4a17425ee29	/creativeWorkStatus
StartYear	The year in which the creation process was started CCR_C-2539_f831f74e-f8ca-4e29-bb02-eb6ca7ea3073	/startDate
CompletionYear	The year in which the creation process was completed CCR_C-2509_3b86afe2-ebde-ba09-8a1c-fe6bdc46a739	/endDate
PublicationDate	The date at which the resource or tool/service was published *i.e.* announced to the public CCR_C-2538_8b697452-7ef3-9fce-ccf9-a7f344f11317	/datePublished
LastUpdate	The date of the last update CCR_C-2526_979ac535-eaa5-5e59-3cad-51c450234698	/dateModified
TimeCoverage	The time period that the content of a resource is about CCR_C-2502_747eb0cd-03e9-cffb-34cc-d0c8c77e4c5a	/temporalCoverage
LegalOwner	The person or institution who/which holds (all) rights to the resource CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb	/copyrightHolder
Genre	The conventionalized discourse or text types of the content of the resource based on extra-linguistic and internal linguistic criteria CCR_C-2470_d191f2b2-6339-f031-b534-70d526b28357	/genre
FieldOfResearch	Indication of the linguistic field for assigning a resource type to its linguistic context. CCR_C-3796_e89bb008-3e2e-1f70-afa5-e506a6c12683	/about

Table 1: Exemplary Mapping from CCR vocabulary to schema.org vocabulary.

```
<Mappings>
  <Schema.org Type (e.g., DataSet)>
    <Context>JSON-LD Context</Context>
    <Profiles>
      <CMD_Profile_Name>CMD Profile identifier</CMD_Profile_Name>
    </Profiles>
    <Mapping>
      <Property>
        <concept>URL</concept>
        <pattern>XPath</pattern>
        <blacklist>CMD Profile identifier</blacklist>
      </Property>
      <Property>
        [...]
      </Property>
    </Mapping>
  </Schema.org Type (e.g., DataSet)>
  <Schema.org Type (e.g., SoftwareApplication)>
    [...]
  </Schema.org Type (e.g., SoftwareApplication)>
</Mappings>

```

Figure 1: General Structure of the Mapping in XML.

that the concepts belong together. It would create a single Person object, with a list of all the given names and family names found in the CMDI.

However, to create a list of Person objects, each with their corresponding first name and family name, it becomes necessary to introduce more context for the mapping system. For this, it supports the `expand` and `expandPattern` elements. With the latter, one can specify the root of a component in the CMD profile via an XPath expression. All further XPath patterns for the properties are then evaluated against this base pattern:

```
<creator expand="true">
  <expand type="Person">
    <expandPattern>XPath base
    pattern</expandPattern>
    <givenName>
      <pattern>XPath</pattern>
    </givenName>
    <familyName>
      <pattern>XPath</pattern>
    </familyName>
  </expand>
</creator>

```

For the mapping to work, the following assumptions are made:

1. For any given overachieving type (*e.g.*, DataSet), all CMD profiles share the same JSON-LD context.
2. Only one XPath expression applies to a given CMD profile. If not, the expressions currently in use are too generic and need to be refined.
3. If a concept is found in a given CMDI, the XPath expressions are not evaluated.

If multiple concepts are found, they will be automatically grouped into a list when converted to JSON-LD. Similarly, if a `lang` attribute is found inside an XML node, it will be automatically converted to a `@language/@value` object. Also, the mapping allows one to nest property/type relations if necessary.
If the concepts and XPath patterns do not yield any result for a given property, it will not be included in the JSON-LD. This avoids null nodes, and it also reduces the size of the resulting files.
Fig. 2 depicts the result of converting a CMDI instance (describing ProFormA, a software application) to JSON-LD. Like the schema.org type DataSet, the type SoftwareApplication is also a sub-type of CreativeWork[6]. The figure, hence, shows many of its properties such as its name, description, genre, funder *etc*. On the right-hand side of the figure, example instantiations of the "licence" and "creator" templates discussed earlier can be seen. Note that both the organisation and the person(s) associated with the creation are given string values as well as Uniform Resource Identifiers (URIs) to refer to their `viaf.org`, `isni.org`, `orcid.org` *etc.* identities.

[6]See `https://schema.org/CreativeWork`.

4. Discussion

The six main profiles we use in our TALAR repository share roughly 80% of their CMDI components (see Sect. 3); those are the metadata that characterise a resource independently of its specific type. The remaining 20% of metadata fields can be used to describe the resource in terms of its specific nature. Lexica can be described in terms of their lexical type and lexical units *etc.*; text corpora can be described in terms of their corpus type, temporal classification *etc.*; and experiments can be described in terms of their experimental paradigm, hypotheses, materials *etc.*
Our work shows that most, if not all, information that is independent of the resource type, can be easily mapped to schema.org vocabulary. The situation is different for terminology that describes the nature of a resource type. Here, no satisfying mapping to schema.org vocabulary is possible. It is this aspect that shows that the CLARIN concept registry has still an important role to play in the CMDI infrastructure.

In fact, we feel that the maintainers of the CCR should fight the well-known proliferation of data categories in the registry by focusing on those metadata fields for which there is a great need in the CLARIN community, and for which existing registries such as schema.org fail to provide adequate vocabulary.
If the CCR maintainers would focus at only providing vocabulary required to adequately describe lexical resources, text and speech corpora, experiments, language-related tools, and other types of linguistic resources, it could throw out the many hundreds of other data categories that are better defined elsewhere. Vocabulary work is by no means trivial and focus is needed to make the CCR a better place to work with.

Our tool converts XML-based CMDI instances that are based on our six CMDI profiles into JSON-LD instances that make use of schema.org terminology. Our TALAR repository of language resources will offer this new JSON-LD-based representation via OAI-PMH harvesting, *complementing* the existing XML-based CMDI export. It shows that the conversion comes with no information loss. Moreover, the new format is understood outside of the CLARIN community, and hence, has the potential to increase the findability of our TALAR resources.
With this positive result, we are investigating whether we should replace our six CMDI profiles whose terminology is semantically grounded with the CCR with new variants where such grounding is achieved via

Figure 2: Metadata Conversion – Fragment.

schema.org. While this step would increase their semantic interoperability across disciplines, it hampers the find-ability of the resources they describe in the CLARIN world. To remedy the situation, we will need to consult with VLO developers so that they extend their facet mapping accordingly. Now, also schema.org terms must be mapped to the dozen criteria (language, format, temporal coverage *etc.*) used for faceted search.

5. Conclusion

In this paper, we have presented an XML-based data structure that describes a mapping between data categories in the CCR and terms in schema.org. A tool has been developed that exploits this representation to map instances of CMDI-based profiles using CCR vocabulary to instances using schema.org terminology. At the same time, the XML-based representation of instances is converted into a JSON-LD-based representation, a format that is understood outside of CLARIN and across communities. We showed that a large majority of data categories in our CMDI profiles can be mapped to schema.org concepts. Those data categories are used to describe those aspects of resources that are independent of their linguistic nature, and for which there exists terminology that is well-defined, widely used, and accepted across communities.

The CLARIN Concept Registry has still an important role to play in the CLARIN community. But we argue that is should focus on defining and providing vocabulary that is not defined elsewhere. For discipline-specific terms, the CCR remains the semantic registry of choice, ensuring that linguistic data is described in an adequate manner. For all other terms, there is usually a discipline-independent, well-maintained, and widely used registry or ontology in place that CMDI practitioners shall use. The advocates of semantic interoperability will thank them.

6. Acknowledgements

Our work has been carried out within the Text+ project[7], funded by the German National Science Foundation (DFG), project reference 460033370.

[7]See `https://www.text-plus.org/en/home/`.

7. Bibliographical References

Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A data category registry- and component-based metadata framework. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta*, pages 43–47. European Language Resources Association.

Clark, J., DeRose, S., et al. (1999). XML Path Language (XPath). Technical report, World Wide Web Consortium. Available at `https://www.w3.org/TR/1999/REC-xpath-19991116/`.

Ďurčo, M. and Windhouwer, M. (2014). The CMD cloud. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'14)*, pages 687–690, Reykjavik, Iceland. European Language Resources Association.

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2):44–51.

ISO 12620. (2009). Terminology and other language and content resources — specification of data categories and management of a data category registry for language resources. International Standard, International Organization for Standardization (ISO), Geneva, 12.

ISO 12620. (2019). Management of terminology resources — data category specifications. International Standard, International Organization for Standardization (ISO), Geneva, 05.

ISO 24622-1. (2015). Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. International Standard, International Organization for Standardization (ISO), Geneva, 01.

ISO 24622-2. (2019). Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language. International Standard, International Organization for Standardization (ISO), Geneva, 07.

Schuurman, I., Windhouwer, M., Ohren, O., and Zeman, D. (2016). CLARIN Concept Registry: The New Semantic Registry. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, Wroclaw, Poland*, Linköping Electronic Conference Proceedings, pages 62–70. Linköping University Electronic Press, Sweden.

Trippel, T. and Zinn, C., (2020). *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, chapter Describing Research Data with CMDI - Challenges to Establish Contact with Linked Open Data. MIT Press. Ed. by A. Pareja-Lora and M. Blume and B. C. Lust and C. Chiarcos.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'12)*, pages 1029–1034. European Language Resources Association.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Windhouwer, M., Indarto, E., and Broeder, D. (2017). Cmd2rdf: Building a bridge from CLARIN to Linked Open Data. In J. Odijk et al., editors, *CLARIN in the Low Countries*, page 95. Ubiquity Press Limited, United Kingdom, December.

Wright, S., Windhouwer, M., Schuurman, I., and Broeder, D. (2014). Segueing from a data category registry to a data concept registry. In *Proceedings of Terminology and Knowledge Engineering 2014*, page 177, Berlin, Germany.

Zinn, C., Hoppermann, C., and Trippel, T. (2012). The ISOcat Registry Reloaded. In E. Simperl, et al., editors, *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 285–299. Springer.

Ďurčo, M. (2013). Smc4lrt - semantic mapping component for language resources and technology. Master's thesis, TU Wien.